STATISTICAL METHODS IN DATA MINING

DR. ALPER VAHAPLAR

Classification in R

- ✓ Decision Trees in R – CART
- ✓ Classification tasks are supervised learning tasks. So we need a training set, and a test set (to check the accuracy of the model)

```
rm(list=ls()) # clear all variables
data(iris) # Loading iris data...
# select 120 rows for training, the rest 30 will be used for
testing
sample = sample(1:nrow(iris),120)
train = iris[sample,]
test  = iris[-sample,]
# Classification Tree Example
library(tree)
# target variable is Species (categorical),
# "." means other variables (y ~ x_i)
model = tree(Species ~., data = train)
# model = tree(Species ~ Sepal.Length + Petal.Length, data =
train) uses 2 variables to model Species
```

Classification in R

✓ Decision Trees in R – CART

```
# model = tree(Species ~., data = train)
summary(model)
plot(model)
text(model)
tahmin = predict(model,test) # view tahmin
tahmin = predict(model,test, type = "class") # view again
table(Tahminler = tahmin, Gerçekler = test$Species)

# Let's look how it works
# Build the model with 2 predictors:
model = tree(Species ~ Petal.Width+Sepal.Width, data=train)
summary(model)
plot(train$Petal.Width, train$Sepal.Width, col=train$Species)
partition.tree(model, label="Species", add=TRUE)
legend("topright",legend=unique(train$Species),
      col=unique(as.numeric(train$Species)), pch=19)
```

Classification in R

- ✓ Decision Trees in R – CART

```
# Regression Tree Example
# this time, target variable is continuous (Sepal.Width)
model2 = tree(Sepal.Width ~.-Species, data=train)
plot(model2)
text(model2)
tahmin = predict(model2,test)
# calculate RMS Error
rmse = sqrt(mean((tahmin-test$Sepal.Width)^2))

# try with different number of predictors
# or with normalized values
```

Classification in R

✓ Decision Trees in R – CART

```
# rpart library

library(rpart)
# Classification Tree in rpart
model3=rpart(Species ~., data=train)
summary(model3)
plot(model3, uniform = T)
text(model3)
printcp(model3) # complexity parameter
plotcp(model3)
tahmin = predict(model3,test, type = "class")
table(Prediction = tahmin, Actual = test$Species)
```

Classification in R

✓ Decision Trees in R – CART

```
# rpart library

library(rpart)
# Regression Tree in rpart
model4=rpart(Sepal.Width~Sepal.Length+Petal.Length+Petal.Width
      , data=train)
plot(model4)
text(model4)
summary(model4)
printcp(model4)
rsq.rpart(model4)
plotcp(model4)
tahmin = predict(model4, test)
rmse = sqrt(mean((tahmin-test$Sepal.Width)^2))
```

Classification in R

✓ Decision Trees in R – C5.0 Tree

```
# C50 library

library(C50)
model5 = C5.0(Species ~., data = train)
model5 = C5.0(Species ~., data = train, rules = T)
summary(model5)
# no tree view if rules=True
plot(model5)
tahmin = predict(model5, test)
table(Prediction = tahmin, Actual = test$Species)
```